BASE RATES VERSUS SAMPLE ACCURACY: COMPETITION FOR CONTROL IN HUMAN MATCHING TO SAMPLE

ADAM S. GOODIE AND EDMUND FANTINO

UNIVERSITY OF GEORGIA AND
UNIVERSITY OF CALIFORNIA, SAN DIEGO

People often place undue weight on specific sources of information (*case cues*) and insufficient weight on more global sources (*base rates*) even when the latter are highly predictive, a phenomenon termed *base-rate neglect*. This phenomenon was first demonstrated with paper-and-pencil tasks, and also occurs in a matching-to-sample procedure in which subjects directly experience case sample (cue) accuracy and base rates, and in which discrete, nonverbal choices are made. In two nonverbal experiments, subjects were exposed to hundreds of trials in which they chose between two response options that were both probabilistically reinforced. In Experiment 1, following one of two possible samples (the unpredictive sample), either response was reinforced with a .5 probability. The other sample (predictive) provided reinforcement for matching on 80% of the trials in one condition but in only 20% of the trials in another condition. Subjects' choices following the unpredictive sample were determined primarily by the contingencies in effect for the predictive sample: If matching was reinforced following the predictive sample, subjects tended to match the unpredictive sample as well; if countermatching the predictive sample was generally reinforced, subjects tended to countermatch the unpredictive sample. These results demonstrate only weak control by base rates. In Experiment 2, base rates and sample accuracy were simultaneously varied in opposite directions to keep one set of conditional probabilities constant. Subjects' choices were determined primarily by the overall accuracy of the sample, again demonstrating only weak control by base rates. The same pattern of choice occurred whether this pattern increased or decreased rate of reinforcement. Together, the results of the two experiments provide a clear empirical demonstration of base-rate neglect.

Key words: matching to sample, base-rate neglect, base rates, sample accuracy, choice, key press, adult humans

---

Recently, the problem of base-rate neglect has captured the attention of behavioral (Rachlin, 1989) and cognitive (Koehler, 1996) approaches to decision making. In base-rate experiments subjects are provided with two sources of information: (a) how often each of two outcomes occurs in a general population (base-rate information), and (b) a cue that bears some relation to the outcome. The subject's task is to select the more likely outcome, or to provide a verbal estimate of the probability of one or both outcomes. Typically, adult humans have been found to underweight the base-rate information and overweight the accuracy of the cue, whether the probabilistic relations are conveyed verbally (Tversky & Kahneman, 1982) or are directly experienced (Goodie & Fantino, 1995, 1996).

A good example of a base-rate problem is given by Eddy (1982), who asked physicians to estimate the probability that a particular woman had breast cancer if she has tested positive in a mammogram, given the following data:

(a) 1% of all women in her reference class have breast cancer.
(b) If a patient has breast cancer, the probability that the mammogram will be positive is 79%.
(c) If a patient does not have cancer, the chance of having a positive mammogram is 9.6%.

The mammogram problem can be solved with reference to the frequency table in Figure 1, which shows expected incidence of positive and negative tests, and breast cancer being present and absent, in a population of 10,000 women (see Gigerenzer & Hoffrage, 1995). It would be more common for the test

Test

	Positive	Negative	
Present	79	21	100
Absent	950	8950	9,900
	1029	8971	10,000

Disease

Fig. 1. Incidence table for 10,000 cases in the mammography problem (Eddy, 1982). Because 1% of all women have breast cancer, .01 × 10,000 = 100 is the marginal sum of the first row, and 9,900 is the sum of the second row. Of the 100 women who have breast cancer, the test would accurately diagnosis 79% of them (100 × .79 = 79) as having breast cancer and would be negative for the other 21% (21). Of the 9,900 women who do not have breast cancer, the test would wrongly identify 9.6% of them (9,900 × .096 = 950) as having the disease, and would correctly be negative for the other 90.4% (8,950). In 10,000 screenings, then, 79 + 950 = 1,029 would come back positive, and of those, 79 would actually be associated with breast cancer. Hence given that the test is positive, there is a 79/950 = 7.7% chance that the woman in fact has breast cancer.

to incorrectly identify a healthy woman as having the disease (.096 × .99 × 10,000 = 950 of the 10,000 cases) than for a woman with the disease to be correctly identified (.79 × .01 × 10,000 = 79 of the 10,000 cases). Thus, given that the test is positive (which happens 79 + 950 = 1,029 times out of 10,000 cases), the probability that the woman actually has breast cancer is 79/1,029 = 7.7%, the correct answer to the problem. However, Eddy (1982) found that few physicians arrived at this answer to the problem.

The mammogram problem is one example of reasoning according to Bayes's theorem, which has been widely adopted as normative (Tversky & Kahneman, 1982). The "odds form" of Bayes's theorem states that

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio.} \quad (1)$$

The posterior odds of a situation is the ratio of the probabilities of two exclusive states of the world (or hypotheses, ordinarily indicated as H_1 and H_2) after some new information (or data, ordinarily indicated as D) has become known. Thus,

$$\text{posterior odds} = \frac{p(H_1|D)}{p(H_2|D)}, \quad (2)$$

where $p(H_1|D)$ is read "the probability of H_1, given D." In the mammogram problem, the hypotheses are that a particular woman has or does not have breast cancer, and the datum is the positive test result.

The prior odds is the ratio of the same probabilities before the new information arrived; thus,

$$\text{prior odds} = \frac{p(H_1)}{p(H_2)}. \quad (3)$$

If we take the presence of breast cancer to be H_1 and its absence to be H_2, then $p(H_1) = 1\%$ and $p(H_2) = 99\%$. The prior odds are .01/.99 = 0.0101.

The likelihood ratio is the ratio of the probabilities that the new information (D) would occur, given that each of the two possible states of the world was in effect; thus,

$$\text{likelihood ratio} = \frac{p(D|H_1)}{p(D|H_2)}. \quad (4)$$

In the mammogram problem, $p(D|H_1) = 79\%$ and $p(D|H_2) = 9.6\%$, so that the likelihood ratio is .79/.096 = 8.23. Substituting the definitions in Equations 2 through 4 into Equation 1,

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(H_1)}{p(H_2)} \times \frac{p(D|H_1)}{p(D|H_2)}. \quad (5)$$

Inserting the prior odds and likelihood ratio in Equation 5 gives posterior odds of 0.0101 × 8.23 = .083. Because $p(H_1|D) + p(H_2|D) = 1$, we can solve for either one, and solving for $p(H_1|D)$ gives the same 7.7% as was derived using the incidence table.

Eddy (1982) and Casscells, Schoenberger, and Graboys (1978) found that physicians of-
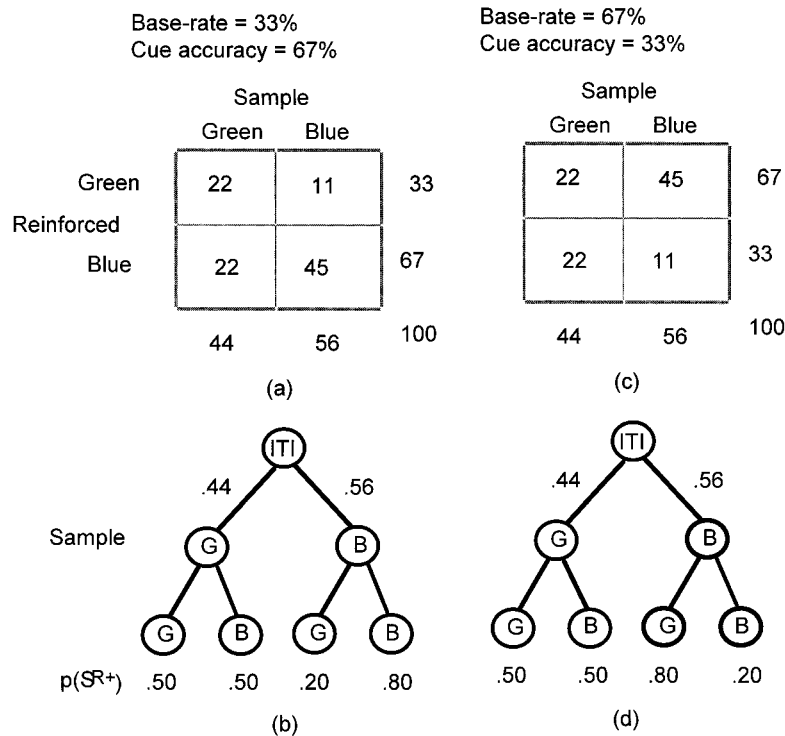
Fig. 2. Incidence tables and procedure (panels a and b) for one condition within an experiment previously conducted (Goodie & Fantino, 1995, Experiment 1), which was replicated in the present Experiment 1. The incidence table and procedure for the other group in Experiment 1 are also given (panels c and d). Following a probabilistically presented sample, each alternative choice is reinforced with the probability given. The groups differ only in the contingencies of reinforcement for choices following a green cue.

ten reported under these circumstances that a woman has a 79% chance of having breast cancer, apparently confusing $p(D|H)$ for $p(H|D)$. More generally, experts and laypeople alike seem to report posterior odds that are too close to the likelihood ratio and not sufficiently affected by prior probabilities. Because many prior probabilities are global base rates comparable to the 1% rate of breast cancer in the population (but see McKenzie, 1994), this phenomenon has been termed *base-rate neglect* (Bar-Hillel, 1980), and has proven to be quite robust (see Koehler, 1996, for a recent review). People respond to a multitude of Bayesian questions by holding the accuracy of a case-specific cue ($p[D|H]$) to closely represent the posterior probability of the event it predicted ($p[H|D]$).

In some studies, subjects are not simply told the probabilities with which various events occur; rather, they are given direct experience with cases, their test results and outcomes, before judging any probabilities.

Some investigators assume that such direct experience with contingencies is equivalent to verbal descriptions of these contingencies. Giving subjects direct contact with probabilistic events confers methodological advantages because an experimenter can more effectively control a subject's history of reinforcement with respect to the repeated events in an experiment than a subject's preexperimental verbal history (Goodie, 1997b).

A literature has emerged to suggest that base-rate neglect does in fact prevail under conditions of direct experience (Edgell, Roe, & Dodd, 1996; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Goodie & Fantino, 1995, 1996; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1990). The basic procedure of one condition from such an experiment (Goodie & Fantino, 1995, Experiment 1) is reproduced in the left side of Figure 2, along with the incidence table from which it derives. It employs a base rate of 33% reinforcement for choosing green and a sam-

ple accuracy of 67%. (Henceforward, each condition will be identified by its base rate and sample accuracy, separated by a slash. Thus, a group experiencing a base rate of 33% and a sample accuracy of 67% is referred to as 33%/67%.) In discussing outcome events, for the sake of clarity we adopt the convention of referring to green as the color whose selection has the greater base rate of being correct. In actuality, though, this variable has always been counterbalanced between subjects so that for half of all subjects, selecting blue is correct more often. This is also true of the experiments conducted in the present study. Because the incidence table in Figure 2a dictates that blue appears as the sample $11 + 45 = 56$ times out of every 100, blue appeared as the sample 56% of the time. When the sample was blue, the reinforced choice was blue 45 of those 56 times; thus, choosing blue following a blue sample was reinforced with $45/56 = 80\%$ probability, and choosing green following a blue sample was reinforced with $11/56 = 20\%$ probability. By identical reasoning, green appeared as the sample on the 44% of trials when blue did not, and when it did, a green choice was reinforced with 50% probability, and a blue choice was reinforced with 50% probability.

Figure 2b depicts the procedure in terms of a delayed matching-to-sample procedure in which an intertrial interval (ITI) is followed by a green or blue sample and then by a choice between green and blue comparisons. The procedure differs from more typical matching-to-sample procedures in that the two samples need not always be equiprobable and a match is not always reinforced. Another way of stating this is to say that the sample accuracy is less than 1.0.

Three sorts of predictions might be made concerning performance under these conditions. The first is that subjects will choose in a way that maximizes reinforcement, namely indifference following green samples and exclusive matching of blue samples. The second prediction in this setting is base-rate neglect, where choice proportions adhere more closely to the global sample accuracy than to the conditional probabilities associated with each sample.

The third prediction is probability matching. A rich literature on the probability learning paradigm exists (Humphreys, 1939; Myers, 1976), wherein subjects are exposed to many unpredictable events to observe how such experience influences their predictions of future events. In general, people in such situations predict the events in proportion to their rates of occurrence (Myers, 1976), a phenomenon termed *probability matching,* which is typically viewed as an error on the subject's part. Each time a subject predicts the less likely event, his or her probability of being "correct" (and procuring whatever reinforcer comes with being correct) declines, reducing the expected payoff. Exclusively predicting the more common outcome would thus maximize reinforcement. Interestingly, pigeons do not exhibit probability matching, but rather maximize under comparable circumstances (e.g., Herrnstein & Loveland, 1975). Moreover, pigeons do not exhibit base-rate neglect. Hartl and Fantino (1996) studied pigeons in a matching-to-sample procedure analogous to those of Goodie and Fantino (1995). Their pigeons typically responded optimally, in terms of maximizing rate of reinforcement, with choice controlled by the sample when it was more discriminative of reinforcement than the base rates, and by base rates when they were more discriminative of reinforcement schedules than the sample. A number of studies have added to the probability learning design by employing a conditional cue that signaled which of two sets of probabilities was in effect. Subjects typically matched the posterior probabilities associated with each cue (see Castellan, 1977, for a review).

In the setting depicted in Figure 2, Goodie and Fantino's (1995) subjects (college students) matched the green sample considerably more than half the time. Because reinforcement is equiprobable given either choice, this fact has no normative implications. However, another group in the same experiment experienced a different sample accuracy, such that matching a blue sample was reinforced 33% of the time, and countermatching blue (i.e., selecting green following a blue sample) was reinforced 67% of the time. Subjects in this group showed a suboptimal pattern of responding, matching the blue sample 56% of the time on average, deviating from both optimization and probability matching accounts. This cost of such deviation was compounded

in another experiment by rewarding correct selections with money, and base-rate neglect persisted (Goodie & Fantino, 1995, Experiment 2). Our experiments (Goodie & Fantino, 1995) differed from those described in the probability learning literature by using cues and outcomes that are identical to each other, and this difference likely accounts for our different data patterns. In support of this hypothesis, follow-up studies (Goodie, 1997a, Experiment 1; Goodie & Fantino, 1996) employed lines as samples that varied on the dimension of orientation, in advance of choice options that varied in color, and probability matching was observed, suggesting that line orientation had acquired appropriate discriminative control.

Surprisingly, this and other studies that demonstrate base-rate neglect under direct experience have not manipulated the base rates that, the researchers conclude, subjects neglect. The inference is instead drawn from data points that deviate from an optimal standard in a direction consistent with base-rate neglect. However, to demonstrate that base rates are in fact neglected, it is desirable to examine behavior in relation to two or more base rates in an experiment and to show that behavior is insufficiently altered thereby.

Two experiments were conducted in an attempt to measure the relative influence of base rates and sample accuracy on choice in a matching-to-sample context. When one manipulates base rates in isolation, however, this inevitably results in comparable changes in the posterior probabilities (see Equation 1), making it difficult or impossible to determine whether the controlling factor in changed behavior is the base rates or the posterior probabilities. However, if one manipulates both base rates and sample accuracy simultaneously, increasing one while decreasing the other, it is possible to balance them so that the posterior probability remains constant. Of the three terms—base rates, sample accuracy, and posterior probability—one increases, one decreases, and one stays constant. The existence and direction of any change in predictive behavior can uniquely identify the term to which behavior is anchored.

EXPERIMENT 1

Two groups of subjects were studied: One experienced conditions of a 67% base rate in favor of green and 33% accurate sample (67%/33%); the other experienced a 33% base rate and a 67% accurate sample (33%/67%). These conditions yield contingencies of reinforcement depicted in panels b and d of Figure 2. Note that in both conditions, the green sample appears on 44% of all trials and is associated with a 50% probability of reinforcement, whichever choice is made. Matching the blue sample is associated with a 20% probability of reinforcement in the 67%/33% condition, but with 80% reinforcement in the 33%/67% condition. In a condition equivalent to the 33%/67% condition, Goodie and Fantino (1995) demonstrated predominant matching to the green sample. The present experiment sought to replicate and extend this effect by asking how often the 67%/33% group would match the green sample.

Three outcomes are possible. First, subjects' choices may be primarily controlled by the rate at which matching the sample is reinforced. Matching the sample is, on the whole, reinforced less frequently for the 67%/33% group than for the 33%/67% group. This hypothesis would therefore be supported if subjects in the 33%/67% group matched the green sample more often than did those in the 67%/33% group. In the past (Goodie & Fantino, 1995, 1996), the 33%/67% group has been observed to match the green sample well over half the time. Although consistent with an interpretation based on preponderant control by overall sample accuracy, more conclusive support would come from comparing the matching frequencies with those of subjects exposed to a 67%/33% arrangement.

A second possibility is that choice is controlled by the relative rates at which choosing the two colors are reinforced, regardless of the sample. Choosing blue is reinforced less often for the 67%/33% group than for the 33%/67% group, so the 67%/33% subjects might be expected to choose blue less often overall, and therefore choose green more, particularly following a green sample. Such an outcome would suggest predominant control by base rates. Although this outcome seemed unlikely in light of all prior research on humans, it would be consistent with previous experiments under comparable conditions with pigeons (Hartl & Fantino, 1996).

These possibilities are in conflict. The first predicts that the 33%/67% group will match the green sample more often than the 67%/33% group will, whereas the second predicts that the 33%/67% groups will match the green sample less often than the 67%/33% group will. These hypotheses also represent predominance in the competition for control of choice by either base rates (the second possibility) or sample accuracy (the first possibility).

The third possibility is that performance would reflect the combined influence of base rates and sample accuracy in a manner consistent with Bayes's theorem. This kind of control would result in probability matching based on the conditional probabilities of reinforcement for each choice option in the presence of each sample. Because green samples in both experimental conditions are associated with the same conditional probabilities of reinforcement, varying the contingencies associated with the blue sample would have no effect on matching of the green sample.

Method

Subjects and Apparatus

Twelve undergraduates at the University of California, San Diego participated as subjects, receiving credit for lower division psychology courses in compensation. The study was approved by the University Institutional Review Board.

Experimental sessions were conducted in a room (3 m by 2.1 m) with white walls that were bare except for a blank chalkboard on the front wall. A desk and table placed along the front wall and a chair in front of the desk furnished the room; a ventilation system reduced external noise. The only window was a one-way mirror in the door, which was near the back of the room. The room was illuminated.

All experimental events were controlled and recorded by a desktop computer with an Intel 386 processor, VGA graphics, and a color monitor. Responses were made on the computer's keyboard. The keyboard sat in front of the CPU and monitor.

Procedure

Stimuli and task. An instruction form told subjects:

In this experiment, you will periodically see two colored shapes, and it is your task to pick the correct one by pressing "d" to pick the one on the left, or "k" to pick the one on the right. You will receive a clue beforehand to help you pick the correct side. Please don't use any outside tools, such as a pencil and paper, to help you remember what you saw.

You will earn one point for each correct answer. The average score is about 120, and a score of 150 or better would place you in the top 20%. Please keep playing for the whole session and try to earn as many points as possible.

The benchmarks provided in the second paragraph were intended to be lower than the true 50 and 80 percentile scores, so that almost all subjects would avoid any aversive effects of being "below average." In fact, the "top 20%" score would have been achieved by random choice. Instructions were read aloud by the experimenter while the subject held and could read a separate copy. The subject's copy was left on the desk during all sessions, and no instructions were given regarding future reference to it. Points had no extrinsic value.

A trial consisted of a 2-s ITI followed by a sample, a choice array, and feedback. The instruction, "Press any key to continue . . .," appeared beneath the sample; pressing any key removed the sample and initiated a 2-s delay. The choice array consisted of a blue rectangle and a green rectangle, displayed side by side on the monitor, counterbalanced by side across trials. Subjects pressed d or k on the keyboard to choose between the two figures, d for the figure on the left and k for the figure on the right. A single press immediately produced either the message "That is correct! You now have [cumulative total] points," or "Sorry, that is incorrect. You still have [cumulative total] points." "Press any key to continue . . ." appeared below the feedback message. Pressing a key removed both messages and initiated a 2-s ITI. Figure 3 presents a detailed schematic diagram of the progression of each trial for subjects in the 33%/67% group. The other group differed only in the probabilities associated with the outcomes.

Subjects were randomly assigned to two groups, termed 67%/33% and 33%/67% after the base rates and global sample accura-
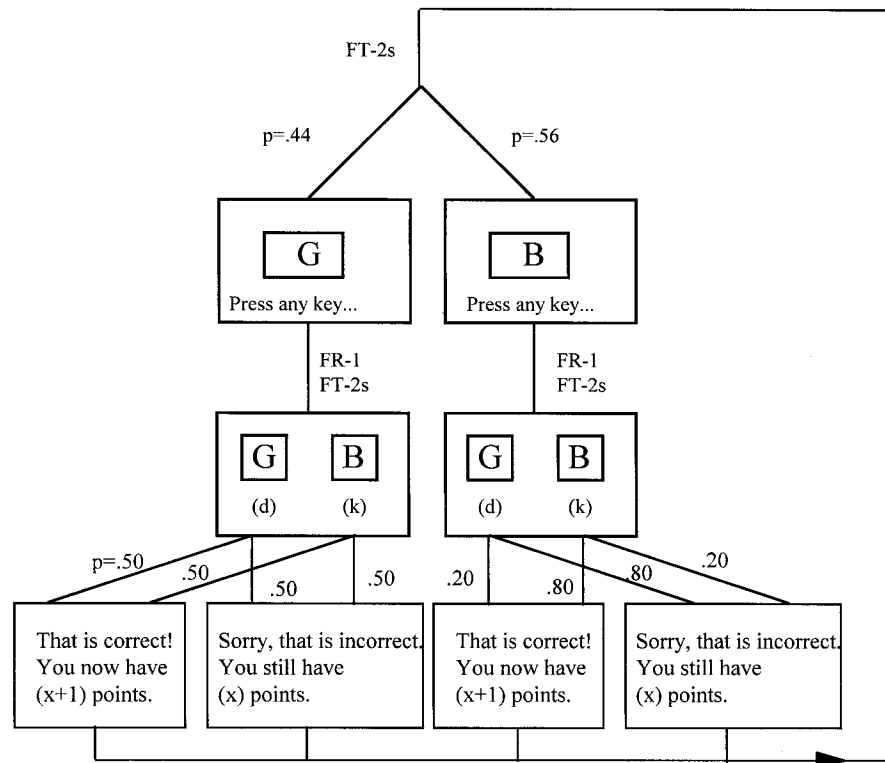
Fig. 3.   Detailed delayed matching-to-sample procedure for the 33%/67% group in Experiment 1. Following a 2-s ITI, a green sample was presented with 44% probability, or else a blue sample was presented. Responding on a fixed-ratio 1 schedule removed the sample and initiated a 2-s retention interval, which was followed by the choice phase. If a subject matched a green sample, reinforcement was presented with 50% probability, or else nonreinforcement was presented. If a subject countermatched a green sample, the probability of reinforcement was still 50%. If a subject matched a blue sample, reinforcement was presented with 80% probability, or else nonreinforcement was presented. If a subject countermatched a blue sample, reinforcement was presented with 20% probability, or else nonreinforcement was presented. The 67%/33% group differed only in the probabilities with which matching and countermatching of the blue sample were reinforced.

cies they encountered. The probabilistic contingencies of reinforcement created by these values are depicted in Figure 2. The color with the greater base rate was counterbalanced between subjects.

Sessions lasted 300 trials or 50 min, whichever passed first. At the end of a session, a message appeared on the computer screen: "The session has now ended." The experimenter could see this message through the one-way mirror, and entered the room to release the subject from the experimental setting. Each subject completed two sessions.

RESULTS AND DISCUSSION

Individual subjects' choice proportions were computed over blocks of 60 trials, and the results are presented in Figure 4. Over-whelmingly, subjects in the 33%/67% group matched green samples more often than did subjects in the 67%/33% group. A t test performed on mean matching proportion over the final four blocks of training was significant, $t(10,$ one-tailed$)$ $=$ 5.53; $p < .05$, and indeed the data reveal no overlap between groups: Every member of the 33%/67% group matched green more often than every member of the 67%/33% group, both over all trials and in the final four blocks. In addition, every subject in the 33%/67% group matched the green sample more than the 50% dictated by probability matching, over all trials and over the final four blocks. Every subject in the 67%/33% group matched the green sample less than 50% of the time overall, and 5 of the 6 did so over the final four
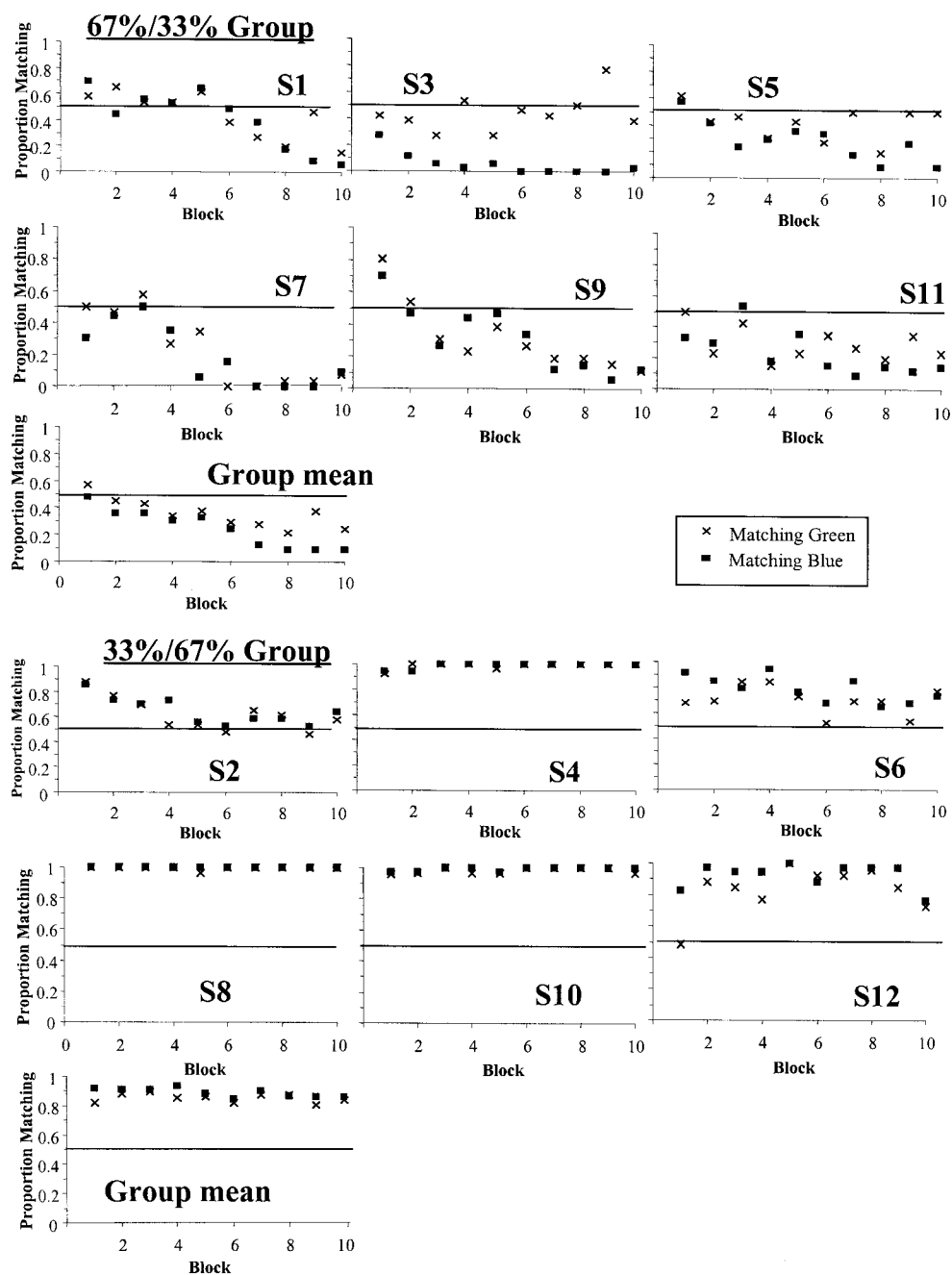
Fig. 4. The results of Experiment 1. Proportions matching green and blue samples are presented for each block of 60 trials. Subjects in the 33%/67% group all matched green samples more often than subjects in the 67%/33% group did, even though they were associated with identical contingencies of reinforcement.

blocks. (S3 matched the green sample 52% of the time.) The mean difference between the groups was 57 percentage points in the final four blocks (85% vs. 28%).

These results suggest that responding was affected more by sample accuracy and less by base rates than Bayes' theorem would suggest. That subjects in the 67%/33% group

did not consistently match blue samples also suggests that the preponderant matching of green samples in prior experiments (e.g., Goodie & Fantino, 1995) was not due primarily to preexperimental bias to match colors.

It cannot be concluded on the basis of these results, however, that there is no bias to match colors. The results from both blue and green samples are not symmetrical around the 50% mark, as one would expect if there were no bias, but rather there is a trend toward matching. Subjects in the 33%/67% condition overshot the probability associated with a green sample by 35 percentage points overall (35 in the final four blocks), but those in the 67%/33% group undershot by only 14 percentage points overall (22 in the final four blocks). Likewise, when faced with an 80% probability of reinforcement for matching blue, subjects in the 33%/67% group overshot this probability by nine percentage points overall; but when faced with an 80% probability of reinforcement for countermatching blue samples, those in the 67%/33% group undershot this mark by five percentage points. Interestingly, this difference largely dissipated in the final four blocks: Subjects in the 33%/67% group overshot the 80% probability by seven percentage points, but the 67%/33% group overshot by 10 percentage points. This dissipation could be due to joint effects of random variation and ceiling effects at stability.

EXPERIMENT 2

Experiment 2 employed three groups of subjects, each of which experienced one of the following base rate/sample accuracy combinations: 60%/70%, 70%/60%, and 40%/30%, as depicted in Figure 5. There are two comparisons of note to be made. The first is between the 60%/70% and 70%/60% groups on rates of matching the green sample. When the base rate of reinforced green trials increases from 60% to 70% and the accuracy of the sample simultaneously decreases from 70% to 60%, these effects exactly counteract each other in Equation 5. The posterior probability of reinforcement for matching a green sample is therefore equal for the two groups (Figure 5a and b). We may ask whether the base rate or the sample accuracy, which again
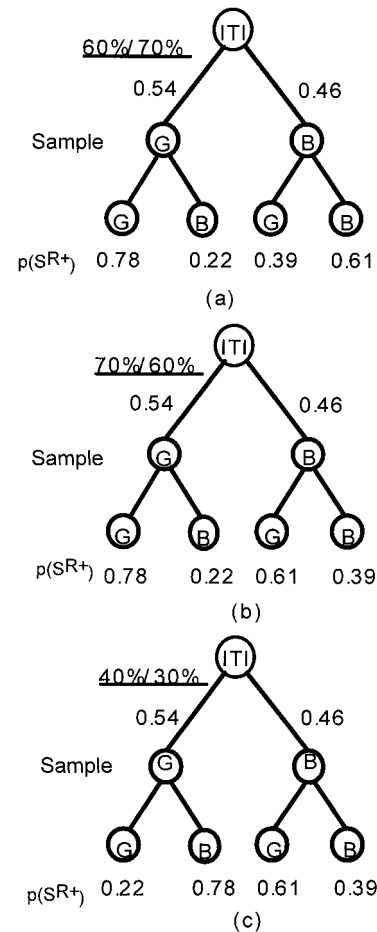

Fig. 5. Incidence tables and procedures for the three groups in Experiment 2. (a) The 60%/70% group was expected to match both samples most of the time, which is not considered base-rate neglect because the contingencies favor preference for matching. (b) Despite no change in the contingencies associated with the green sample, the 70%/60% group was expected to match less than the 60%/70% group, an effect that would reduce obtained reinforcement. Matching of the blue cue was expected to decrease, which would increase reinforcement. (c) Despite no change in the contingencies associated with the blue cue, the 40%/30% group was expected to match it less than the 70%/60% group. This would increase reinforcement.

varied in opposite directions, has primary control over choice. (Also note that the proportion of trials on which each sample appeared remained constant, removing one class of explanations for any observed effects.) For the comparison between the 70%/60% and 40%/30% groups, we consider the effect of the differences on rates of matching

the blue sample, which is associated with the same contingencies of reinforcement. The 70% base rate experienced by subjects in the 70%/60% group is the base rate of reinforced green trials, making the base rate of reinforced blue trials 30%. Similarly, subjects in the 40%/30% group have a 40% base rate of reinforced green trials and therefore a 60% base rate of reinforced blue trials. Thus, for trials with a blue sample, the relevant base rate has increased (between the groups) from 30% to 60%, and the sample accuracy has declined from 60% to 30%. These contradictory tendencies exactly counteract each other, allowing for the equal posterior probabilities depicted in Figure 5b and c.

The predictions of two versions of the hypothesis of base-rate neglect may be considered here. According to the stronger version, only sample accuracy influences choice. Little difference in choice proportions following a blue sample should be observed between the 60%/70% and 70%/60% groups, because the sample accuracy differs by only 10% between these two conditions (70% vs. 60%). This is inconsistent with both probability matching, which implies a difference of 22 percentage points (the difference in reinforcement rates for matching blue in Figure 5a and b; 61% vs. 39%), and optimization, which requires a difference of 100% (choosing blue exclusively following a blue sample, Figure 5a, and choosing green exclusively following a blue sample, Figure 5b). In the condition depicted in Figure 5a, subjects maximize reinforcement by always choosing blue following a blue sample; in the condition depicted in Figure 5b, they maximize reinforcement by always choosing green following a blue sample. Also, although the conditional probabilities associated with the blue sample are the same in the 70%/60% and 40%/30% groups (see Figure 5b and c), the difference between the groups' global sample accuracy is three times as great as between the 60%/70% and 70%/60% groups (30% difference vs. 10% difference in sample accuracy summed over all trials). The strong hypothesis therefore predicts a larger shift in choice proportions between the second two groups.

The strength of these predictions derives from the assumption that base rates have no effect on choice, an assumption that is generally stronger than the data warrant (Koeh-ler, 1996). The weaker version of the hypothesis makes claims only about responding to samples associated with equal conditional probabilities of reinforcement for matching, claiming that those groups experiencing greater overall sample accuracy will match these samples more than groups with lesser overall sample accuracy, despite their equal conditional probabilities. This is the version of the hypothesis employed in Experiment 1. Hence, the prediction was made that the 60%/70% group would match the green sample more than the 70%/60% group because of its higher sample accuracy, despite identical probabilities of reinforcement associated with the green sample and a lower base rate of green outcomes. For the same reason, it was also predicted that the 70%/60% group would match the blue sample more than the 40%/30% group.

METHOD

Subjects. Eighteen undergraduates at the University of California, San Diego participated as subjects, receiving credit for lower division psychology courses in compensation.

Procedure. The apparatus, stimuli, trial structure, and instructions were as they had been in Experiment 1. Subjects completed one session of 300 trials or 50 min, whichever passed first. Sessions were terminated in the same manner as they were in Experiment 1.

Three groups of subjects were studied, termed 60%/70%, 70%/60%, and 40%/30% after the base rates and global sample accuracies they encountered. The probabilistic contingencies of reinforcement created by these values are depicted in Figure 5. Subjects were randomly assigned to the two groups, and the color with the greater base rate was counterbalanced between subjects.

RESULTS AND DISCUSSION

The results are presented in Figure 6, with individual data presented in Table 1. The weaker hypothesis predicted that subjects in the 60%/70% group would match green samples more than subjects in the 70%/60% group, and that subjects in the 70%/60% group would match blue samples more than subjects in the 40%/30% group. The stronger hypothesis made these predictions but also predicted that although subjects in the 60%/70% group would match blue samples
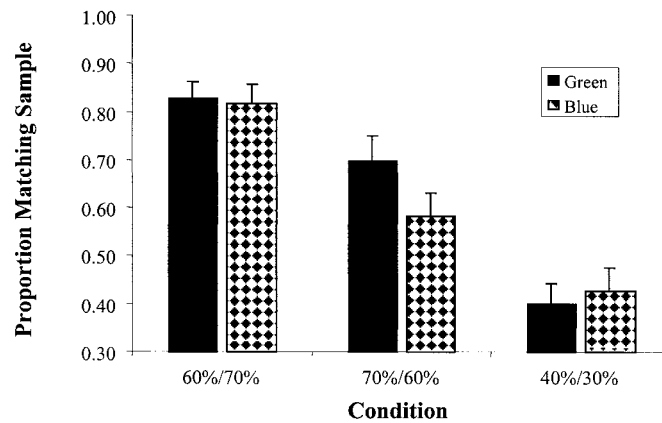
Fig. 6. Overall matching proportions in the three conditions of Experiment 2. The predictions of the weaker base-rate neglect hypothesis were confirmed, but not those of the stronger hypothesis. Error bars represent standard error of the mean.

more than subjects in the 70%/60% group, this effect would be relatively small.

Both of the effects predicted by the weaker hypothesis were evident at the group level: Matching declined with overall sample accuracy, even as the base rate of the color in question increased, and the conditional Bayesian probabilities associated with the sample were unchanged. In the case of the green comparison, the base rate of green events increased (between the groups) from 60% to 70%. In the case of the blue comparison, the base rate of green events declined from 70% to 40%, so that the base rate of

Table 1

Choice proportions of each subject in each experimental group matching each sample in Experiment 2.

	Proportion matching green					Proportion matching blue				
	Block					Block				
Subject	1	2	3	4	5	1	2	3	4	5
60%/70% group										
3	.563	.938	.781	.688	.719	.593	.893	.786	.786	.500
6	.844	.844	.969	.969	.969	.741	1.000	.929	1.000	.929
9	.938	.875	.719	.688	.688	.926	.857	.786	.893	.857
12	.469	.563	.906	.906	.938	.556	.500	.679	.821	.857
15	.844	.750	.719	.906	.781	.852	.893	.750	.786	.714
18	.844	.969	1.000	1.000	1.000	.704	.964	1.000	.929	1.000
70%/60% group										
1	.625	.719	.781	.750	.875	.667	.786	.893	.643	.786
4	.844	.813	.813	.875	.563	.593	.607	.679	.714	.536
7	.531	.500	.469	.625	.375	.519	.464	.714	.464	.393
10	.563	.875	.781	.906	.688	.667	.393	.643	.214	.286
13	.875	.938	.750	.750	.813	.741	.679	.679	.571	.643
16	.719	.531	.469	.531	.563	.333	.464	.643	.464	.571
40%/30 group										
2	.313	.688	.344	.250	.594	.444	.500	.607	.393	.571
5	.469	.375	.281	.406	.438	.630	.250	.536	.393	.357
8	.375	.094	.063	.000	.156	.481	.036	.036	.143	.536
11	.250	.469	.375	.438	.469	.259	.393	.179	.643	.357
14	.469	.344	.406	.281	.250	.407	.571	.464	.714	.857
17	.563	.594	.188	.219	.313	.185	.714	.393	.357	.357

blue events increased from 30% to 60%. Both effects were statistically significant over the entire span of training; for the green comparison, $t(10$, one-tailed$) = 1.96$, $p < .05$; for the blue comparison, $t(10$, one-tailed$) = 2.23$, $p < .05$. However, the support for base-rate neglect was less evident at the level of individual subjects, where there was considerable variability within and between subjects and overlap between groups (see Table 1). In particular, the greater within-subject variability evident in the comparison of matching proportions associated with the blue sample in the 70%/60% and 40%/30% groups suggests that overly strong conclusions should not be drawn from this comparison.

The strong version of the base-rate neglect hypothesis was not supported. Despite encountering samples that were only 10% more accurate, the 60%/70% group matched the blue sample considerably more than 70%/60% group did (average choice proportions of .82 vs. .58, a difference of .24). This difference was statistically significant, $t(10) = 3.65$, $p < .05$, and was greater than that between the 70%/60% and 40%/30% groups (.58 vs. .43, a difference of .15). The strong version of the base-rate neglect hypothesis predicted that the latter difference would be greater because it was associated with a greater difference in overall sample accuracy. The findings represent more evidence that base-rate neglect is not absolute. However, they also represent further evidence that base rates influence choice less than would be expected within a Bayesian framework.

GENERAL DISCUSSION

These two experiments compared choice across pairs of conditions in which subjects experienced the same conditional probabilities of reinforcement for matching one of the colors (e.g., green), but for which base rates and overall sample accuracies both differed. By varying base rate and sample accuracy in opposite directions (i.e., by increasing the base rate of one color while lowering sample accuracy), the conditional probability of reinforcement was held constant. A major question was whether choice of a color would remain constant or, if not, would it deviate from constancy in the direction of the change in base rates or toward the sample accuracy? In

other words, would base rates or sample accuracy be more effective in controlling matching-to-sample responding? In all cases, choice was controlled primarily by sample accuracy, consistent with the phenomenon of base-rate neglect. This effect was most evident in Experiment 1. Although the data in Experiment 2 tended in the same direction, within- and between-subjects variability was considerably greater.

Although these results are consistent with previous reports of base-rate neglect, it should be emphasized that the results also demonstrate some control over behavior by the base rates. Base rates did influence choice: Subjects in all groups made the more reinforced match more often than they made the less reinforced match. Because overall sample accuracy was distributed equally to both reinforced colors, if overall sample accuracy were the sole determinant of choice, the blue sample would have been matched as often as the green sample. That this did not occur shows some degree of control by the base rates.

The present studies demonstrate that a matching-to-sample procedure can be adapted to capture the essential conditions that define the base-rate problem, the subject of scores of studies and reviews (e.g., Koehler, 1996). These prior studies have typically employed a procedure in which subjects encounter one or a few tasks involving verbal stimuli and hypothetical outcomes, and in which verbal responses are recorded. Perhaps with repeated reinforced trials, base-rate neglect would dissipate and behavior would become more optimal, at least reaching the levels found in probability learning experiments discussed earlier. Repeated trials with questionnaires have yielded mixed results. Fischhoff, Slovic, and Lichtenstein (1979) and Birnbaum and Mellers (1983) found that with multiple verbal problems employing various base rates, subjects' probability estimates varied in accordance with the base rates. Lindeman, Van Den Brink, and Hoogstraten (1988) replicated these results but found no carryover to a novel problem. Base-rate neglect prevailed even in subjects who had been informed of correct responses in previous phases, suggesting that Bayesian reasoning per se may not have been strengthened in the training trials. The matching-to-sample pro-

cedure provides a second way of assessing the robustness of base-rate neglect to repeated encounters. In the present experiments, base-rate neglect did not dissipate substantially with repeated trials.

A methodological limitation of the present study concerns the use of a between-subjects design and the degree to which experimental control was achieved at the level of individual subjects. This was a more substantial limitation in Experiment 2, where variability was considerable and there was overlap between groups. However, the clear between-groups differences in Experiment 1, coupled with the low within-group variability, suggest that the effects were real at the individual-subject level.

The Difference Between Pigeons and People

The present results are consistent with our previous data and conclusions based on adult humans as subjects (Goodie & Fantino, 1995; 1996), but stand in contrast with those of Hartl and Fantino (1996) with pigeons as subjects. As in the present Experiment 1, Hartl and Fantino arranged conditions in which one sample was predictive (matching was reinforced with 80% probability) and another sample was unpredictive (selecting either of two comparisons was reinforced on half of the trials). Unlike our humans, the pigeons in Hartl and Fantino's experiment tended to countermatch the unpredictive sample, demonstrating predominant control by the base rates of the reinforced stimuli.

One possible explanation is that, unlike adult humans, Hartl and Fantino's (1996) pigeons had not acquired generalized matching as an operant and therefore could not have acquired behavior that depends on the global accuracy of the sample. Wright and his colleagues (Wright, Cook, Rivera, Sands, & Delius, 1988; Wright & Delius, 1994) have identified conditions that give rise to certain kinds of generalized identity matching in pigeons. For example, when samples are drawn from a large set, generalized identity matching prevails (see also Zentall & Hogan, 1978). It is possible that with such prior training, pigeons would, like adult humans, fall prey to base-rate neglect.

Another possibility is that the differences between pigeons and humans are related to differences in procedure. Unlike reinforcers typically used with pigeons, the points we used in the present study had no extrinsic value, but rather had reinforcing effectiveness that was established through instructions. These instructions may have reduced sensitivity to programmed reinforcement contingencies (Baron & Galizio, 1983).

However, the data contain evidence that the points we used were in fact reinforcers. First, consider the rate at which blue samples were matched in Experiment 1. Recall that these samples differed only in how frequently matching was reinforced. For the 33%/67% group, matching was reinforced with 80% probability, and matching was observed 87.3% of the time over the final four blocks; for the 67%/33% group, matching was reinforced with 20% probability and was observed 9.9% of the time over the same period. Thus, the differential reinforcement of matching resulted in a difference of 77 percentage points in choice proportions following the noncrucial blue sample. Second, in all conditions except one (Experiment 2, 70%/60% group, blue sample), choice proportions were in alignment with the probabilities of reinforcement for these choice alternatives. Third, that the present results are in general agreement with some of our prior work using monetary reinforcers suggests that the effects reported here are not limited to instructed reinforcers with no extrinsic value.

These experiments assessed matching-to-sample behavior as a function of sample accuracy and as a function of the frequency of reinforcement associated with the outcome selected (base rates). Although both variables exerted control on subjects' responding, the effect of sample accuracy predominated, resulting in nonoptimal behavior. These results also constitute a behavioral demonstration of what has been termed base-rate neglect. In addition to exploring the determinants of matching-to-sample responding with humans, an aim of the present studies was to employ the behavioral methods of matching to sample to explore a phenomenon typically studied by cognitive and social psychologists, an approach that is increasingly common in the experimental analysis of behavior (e.g., Davison & Tustin, 1978; Donahoe & Palmer, 1989, 1994; Fantino, 1998; Kehoe, 1989; Rachlin, 1989; Rachlin, Logue, Gibbon, & Frankel, 1986; Shull, 1995; Silberberg, Mur-

ray, Christensen, & Asano, 1988; White, 1985). Interestingly, there are also impulses within the field of judgment and decision making to pursue interactions with more analytical and less normative branches of psychology. Barbara Mellers (1996), then the President of the Society for Judgment and Decision Making, warned recently that her field is at "the point of diminishing returns" and suggested broadly that "we need theories of decision making that predict not only errors, biases, and violations of axioms, but also broader themes of psychological and social functioning" (p. 3). There is in all likelihood further room for fruitful interaction between behavior analysts and researchers in judgment and decision making.

REFERENCES

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233.

Baron, A., & Galizio, M. (1983). Instructional control of human operant behavior. *Pscyhological Record, 33,* 495–520.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45,* 792–803.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299,* 999–1001.

Castellan, N. J. (1977). Decision making with multiple probabilistic cues. In N. J. Castellan, D. P. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 117–147). Hillsdale, NJ: Erlbaum.

Davison, M. C., & Tustin, R. D. (1978). The relation between the generalized matching law and signal-detection theory. *Journal of the Experimental Analysis of Behavior, 29,* 331–336.

Donahoe, J. W., & Palmer, D. C. (1989). The interpretation of complex human behavior: Some reactions to *Parallel Distributed Processing,* edited by J. L. McClelland, D. E. Rumelhart, and the PDP Research Group. *Journal of the Experimental Analysis of Behavior, 51,* 399–416.

Donahoe, J. W., & Palmer, D. C. (1994). *Learning and complex behavior.* Boston: Allyn & Bacon.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge: Cambridge University Press.

Edgell, S. E., Roe, R. M., & Dodd, C. H. (1996). Base rates, experience, and the big picture. *Behavioral and Brain Sciences, 19,* 21.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556–571.

Fantino, E. (1998). Behavior analysis and decision making. *Journal of the Experimental Analysis of Behavior, 69,* 355–364.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance, 23,* 339–359.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Goodie, A. S. (1997a). Base-rate neglect under direct experience. *Dissertation Abstracts International, 58* (01), 435 B.

Goodie, A. S. (1997b). Direct experience *is* ecologically valid. *Behavioral and Brain Sciences, 20,* 777–778.

Goodie, A. S., & Fantino, E. (1995). An experientially derived base-rate error in humans. *Psychological Science, 6,* 101–106.

Goodie, A. S., & Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature, 380,* 247–249.

Hartl, J. A., & Fantino, E. (1996). Choice as a function of reinforcement ratios in delayed matching to sample. *Journal of the Experimental Analysis of Behavior, 66,* 231–242.

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching in concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior, 24,* 107–116.

Humphreys, L. G. (1939). Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology, 25,* 294–301.

Kehoe, E. J. (1989). Connectionist models of conditioning: A tutorial. *Journal of the Experimental Analysis of Behavior, 52,* 427–440.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19,* 1–53.

Lindeman, S. T., Van Den Brink, W. P., & Hoogstraten, J. (1988). Effect of feedback on base-rate utilization. *Perceptual and Motor Skills, 67,* 343–350.

McKenzie, C. R. M. (1994). Base rates versus prior beliefs in Bayesian inference. *Psycoloquy* 5 (1) base-rate.6.mckenzie.

Mellers, B. (1996). From the president. *Judgment and Decision Making Newsletter, 15,* 3.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 3, pp. 171–205). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

Rachlin, H. (1989). *Judgment, decision, and choice: A cognitive/behavioral synthesis.* New York: Freeman.

Rachlin, H., Logue, A. W., Gibbon, J., & Frankel, M. (1986). Cognition and behavior in studies of choice. *Psychological Review, 93,* 33–45.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology, 42,* 209–237.

Shull, R. (1995). Interpreting cognitive phenomena: Re-

view of Donahoe and Palmer's *Learning and Complex Behavior*. *Journal of the Experimental Analysis of Behavior, 63,* 347–358.

Silberberg, A., Murray, P., Christensen, J., & Asano, T. (1988). Choice in the repeated-gambles experiment. *Journal of the Experimental Analysis of Behavior, 50,* 187–195.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge: Cambridge University Press.

White, K. G. (1985). Characteristics of forgetting functions in delayed matching to sample. *Journal of the Experimental Analysis of Behavior, 44,* 15–34.

Wright, A. A., Cook, R. G., Rivera, J. J., Sands, S. F., & Delius, J. D. (1988). Concept learning by pigeons: Matching-to-sample with trial-unique video picture stimuli. *Animal Learning & Behavior, 16,* 436–444.

Wright, A. A., & Delius, J. D. (1994). Scratch and match: Pigeons learn matching and oddity with gravel stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 20,* 108–112.

Zentall, T. R., & Hogan, E. (1978). Same/different concept learning in the pigeon: The effect of negative instances and prior adaptation to transfer stimuli. *Journal of the Experimental Analysis of Behavior, 30,* 177–186.